

Diversity and Deliberation in Homogeneous LLM Prediction Polls

Jonas Gebele*, Tristan Till*, Florian Matthes

Technical University of Munich

School of Computation, Information and Technology

Abstract

Multi-agent large language model (LLM) systems increasingly combine model-generated judgments for forecasting and collective decision-making. However, homogeneous agent systems, i.e., those that share a model and evidence, may produce correlated forecasts, and communication may further contract the diversity that remains. We examine these issues in live LLM prediction polls. Across 189 resolved snapshots, mean residual-error correlation was 0.910. Semantic prompt differences were associated with forecast differences, but diversity-aware selection did not detectably outperform prespecified random ensembles. Standard deliberation removed approximately twice as much disagreement benefit as it added through improvements in individual forecasts at the point estimates, with no detectable improvement in aggregate accuracy. In a prospective intervention, a probability-blind summary of forecast-relevant claims retained 93% of initial dispersion, compared with 29% under standard exchange.

1 Introduction

Multi-agent LLM systems aggregate or deliberate over judgments from prompt-conditioned instances. Prediction polls make their collective value testable: agents forecast unresolved events, and realized outcomes permit evaluation of dependence, aggregation, and revision. Yet crowd wisdom depends on complementary errors, not nominal group size or textual role variation (Atanasov et al., 2017).

Homogeneous polls may face two successive bottlenecks. Before interaction, shared model weights and evidence can produce correlated forecasts, limiting the value of additional agents. During interaction, peer probabilities and rationales may contract the remaining disagreement without adding information. We therefore ask: **RQ1:** Does prompt

diversity create independently useful predictive signals in a homogeneous poll? **RQ2:** How does communication transform the diversity that remains?

We examine these questions using outcome-resolved forecasts and a matched prospective communication intervention. Holding the model and retrieved evidence constant separates prompt variation from differences in search results, although single responses cannot fully separate prompt effects from sampling variation.

Our contributions are:

- We characterize forecast dependence and the relationship between semantic prompt variation and forecast differences, and test diversity-aware selection against a prespecified random ensemble.
- Using the Brier decomposition, we distinguish individual error changes from the disagreement benefit lost through convergence during standard deliberation.
- Through matched communication conditions, we compare conformity under peer probabilities, raw rationales, and probability-blind crux sharing, without equating retained dispersion with improved accuracy.

2 Related Work

Forecasting and diversity. Heterogeneous LLM ensembles can approach human-crowd forecasting performance (Schoenegger et al., 2024). Same-model ensembles instead vary reasoning paths or prompts (Wang et al., 2023; Hu et al., 2025), but correlated errors can limit their gains (Aitchison et al., 2026). Persona conditioning does not reliably improve calibration (Iadisernia and Camassa, 2025). Effective rank measures dimensionality (Roy and Vetterli, 2007) and has been applied to LLM outputs (Yang et al., 2026; Patel, 2026); we use it to describe variation.

*These authors contributed equally to this work.

Elicited probabilities. Our forecasts are numerical judgments generated as tokens, compared to other methods focusing on token probabilities. Verbalized confidence can improve calibration in question answering (Tian et al., 2023), but overconfidence is also documented (Groot and Valdenegro-Toro, 2024). Neither establishes calibration for future-event forecasting; Brier scores alone do not isolate calibration or rule out overconfidence.

Deliberation. Debate benefits depend on group composition and communication (Du et al., 2024; Smit et al., 2024). Shared-information forecasters gain little from deliberation, whereas complementary evidence can support useful exchange (Schneider and Schramm, 2025; Li et al., 2026). Studies document premature consensus and diversity collapse (Chen et al., 2026; Wu and Ito, 2025), echoing human-crowd findings (Lorenz et al., 2011). We distinguish individual error changes from disagreement loss and compare communication formats.

3 Study Design and Measures

3.1 Homogeneous Prediction Polls

We collect forecasts before resolution for binary prediction markets sourced from Polymarket in politics, economics, and geopolitics, withholding contemporaneous market prices. Each snapshot has one price-blind research brief shared by all agents, holding model and evidence fixed within a poll. The main experiment uses 30 prompt-conditioned gpt-5.4-mini agents; after resolution and quality filtering, evaluation covers 189 snapshots from 165 markets. A prospective intervention uses eight fixed gpt-5.6-terra agents on 40 frozen, unresolved event clusters. Code, prompts, and settings are available in a public repository.¹

3.2 Experimental Comparisons

Prompt selection. We generate 285 prompts combining three persona prefixes (Appendix B) with distinct reasoning instructions, evaluated with gpt-4.1-mini on 150 development markets resolved after its June 1, 2024 knowledge cutoff. DIPPER’s Fidelity-Adjusted Semantic Volume (FASV) criterion (Hu et al., 2025) selects 30 prompts. FASV combines a prompt-embedding Gram log-determinant, rewarding nonredundant

semantic directions, with weighted development scores; the published selector builds the subset greedily. We compare it with one prespecified random 30-prompt ensemble from the same pool. Independent forecasts are averaged arithmetically.

Standard deliberation. The 30 selected agents are randomly split into two teams of 15. Each reports a probability of a Yes resolution and a one-sentence rationale, then revises after seeing teammates’ anonymized probabilities and rationales. We obtain 7,499 matched revisions across 500 team–market panels; one initial request yielded no valid probability.

Matched communication intervention. Identical frozen forecasts and briefs are branched into five conditions: (1) **self-revision**, without peers; (2) **probability-only**; (3) **rationale-only**; (4) **standard exchange**, with both; and (5) **crux-ledger**, a probability-blind, deduplicated summary of decision-relevant claims without agent identities or source attribution. A *crux* is a claim or assumption whose assessment could change a forecast; the ledger consolidates repeated claims. *Probability-blind crux sharing* is our label for this format, not an established algorithm. Removing numerical probabilities does not remove directional information from the text. All conditions begin from the same initial forecasts and research briefs, isolating the effect of communication content on revision. Unresolved outcomes permit evaluation of revision, not forecasting accuracy.

3.3 Measures

Let m index a snapshot, i an agent, and N the poll or team size. The outcome y_m is 1 for Yes and 0 otherwise; $p_{im} \in [0, 1]$ forecasts Yes, not rationale correctness, and $\bar{p}_m = N^{-1} \sum_i p_{im}$.

Accuracy and aggregation. Under Brier loss, mean-forecast error decomposes as (Krogh and Vedelsby, 1994)

$$(\bar{p}_m - y_m)^2 = \frac{1}{N} \sum_i (p_{im} - y_m)^2 - \frac{1}{N} \sum_i (p_{im} - \bar{p}_m)^2.$$

The terms are mean individual error and the *disagreement benefit* of averaging. This variance term is an exact gain relative to mean individual error. Their changes separate individual improvement

¹<https://github.com/jonasgebele/diversity-deliberation-prediction-polls>

from convergence. For uniform k -agent subsets without replacement, the fraction of the full-poll gain is necessarily $N(k-1)/[k(N-1)]$, when that gain is nonzero. These percentages cannot diagnose independence (Appendix C). Larger pairwise forecast distance increases gain relative to mean individual error, but need not improve absolute forecasting performance when constituent forecasts are worse.

Forecast diversity. We report pairwise forecast and residual-error correlations, and relate semantic embedding distance to forecast-vector root-mean-square distance and residual dissimilarity, $1 - \text{Corr}_m(p_{im} - y_m, p_{jm} - y_m)$. Mantel permutations account for dependence among prompt pairs. As a complementary dimensionality diagnostic, we report entropy effective rank $r_{\text{eff}} = \exp(-\sum_j q_j \log q_j)$, where q_j are the normalized eigenvalues of the forecast covariance matrix (Roy and Vetterli, 2007). This covariance-spectrum measure is near one when variation lies mainly along one direction and reaches N when all eigenvalues are equal and positive. It describes how forecast variance is spread, not the number of independent agents or predictive usefulness (Appendix C).

Revision and social influence. For initial forecast $p_{im}^{(0)}$ and revision $p_{im}^{(c)}$, let $\Delta p_{im}^{(c)} = p_{im}^{(c)} - p_{im}^{(0)}$. We fit standard deliberation and each prospective condition c separately:

$$\Delta p_{im}^{(c)} = \alpha_c + \beta_c(\bar{p}_{-i,m}^{(0)} - p_{im}^{(0)}) + \epsilon_{im}^{(c)},$$

where $\bar{p}_{-i,m}^{(0)}$ averages the other agents in the team or poll. Larger β_c means stronger movement toward peers, not accuracy; for self-revision, this mean is only an analysis covariate, not shown to agents. Dispersion retention is the post/initial within-panel standard-deviation ratio, averaged across markets with matched agents. Percentile bootstrap intervals cluster by market and preserve matched snapshots, but do not account for cross-market event dependence.

4 Results

4.1 Forecast Dependence and Prompt Selection

Across 189 resolved snapshots, the 30-agent mean achieved Brier 0.2108, outperforming both uninformative baselines: the fixed-0.5 forecast (0.2500) and the uniform-random forecast (0.3333). It did

not outperform the prediction-market benchmark (0.1472; Table 1). The expected one-agent Brier score was 0.2289, giving an absolute reduction of approximately 0.0181 for the 30-agent mean (Figure 1, left). The curve’s gain shares are fixed by subset size. Average forecast correlation was 0.728, realized residual correlation was 0.910, and forecast-covariance effective rank was 4.10, indicating that variation was concentrated in a few dominant dimensions.

Semantic prompt distance correlated with forecast distance ($r = 0.573$) and residual dissimilarity ($r = 0.534$); both Mantel $p < .001$. This variation did not yield a detectable selection advantage: DIPPER achieved Brier 0.2108, versus 0.2118 for the prespecified random ensemble (difference 0.0010, 95% CI $[-0.0048, 0.0063]$). Semantic prompt variation was therefore associated with forecast differences, but diversity-aware selection did not show a detectable aggregation advantage in this comparison.

4.2 Standard Deliberation Contracts Diversity

Across 7,499 revisions, agents moved strongly toward their peers ($\beta = 0.805$, 95% CI $[0.779, 0.830]$), and dispersion declined in 499 of 500 panels. At the point estimates, mean individual Brier improved by 0.0048, but disagreement benefit fell by 0.0093. Consequently, aggregate Brier showed no detectable improvement (change $= +0.0045$, 95% CI $[-0.0010, 0.0099]$). Moreover, 91.6% of 226 initially correct agents facing an incorrect peer majority moved toward that majority. This net loss in aggregate accuracy occurs because social influence is largely asymmetrical: rather than superior arguments convincing the group, minority agents holding correct forecasts routinely capitulate to wrong peer majorities (91.6%). Consequently, standard deliberation does not filter out shared model biases, instead homogenizing the crowd around the model’s dominant priors.

4.3 Communication Content Determines Contraction

The matched prospective intervention shows that contraction depends on the information exchanged. Susceptibility was low under self-revision ($\beta = 0.078$) and the crux-ledger condition ($\beta = 0.124$), intermediate under probability-only ($\beta = 0.484$) and rationale-only ($\beta = 0.502$) exchange, and high-

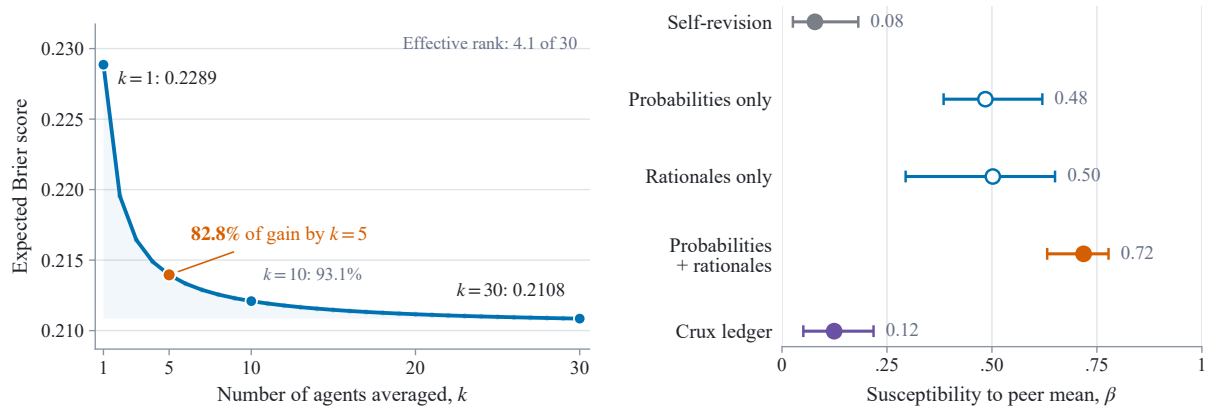


Figure 1: **Forecast aggregation and communication.** Left: expected Brier score of a random k -agent mean, with forecast-covariance effective rank. The labeled gain shares follow from uniform subset averaging, not agent independence. Right: peer susceptibility by revision condition, with 95% market-bootstrap CIs. Probability- and rationale-only conditions are exploratory; prospective outcomes were unresolved at evaluation.

est under standard exchange ($\beta = 0.718$; Figure 1, right).

The crux ledger retained 93% of initial dispersion, similar to self-revision (90%), compared with 63% under rationale-only, 39% under probability-only, and 29% under standard exchange. Although probability-only and rationale-only produced similar susceptibility estimates, rationale-only retained more dispersion, indicating that movement toward peers and total panel contraction capture distinct aspects of revision. By deduplicating claims and withholding both numerical probabilities and author identities, the crux ledger eliminates social majority pressure and numerical anchoring. This forces agents to re-evaluate the factual argument space on its merits rather than converging toward an apparent peer consensus.

5 Discussion

The main poll exhibits strongly correlated residual errors, while semantic prompt distance is associated with forecast variation. Neither this association nor the null comparison with a random ensemble establishes that prompt diversity is generally unhelpful. The relevant test is held-out forecasting value after accounting for constituent accuracy. Subset gain percentages alone cannot identify an optimal crowd size or quantify a penalty attributable to shared model weights.

Standard deliberation reduced disagreement benefit more than it improved mean individual error at the point estimates, with no detectable aggregate improvement. Consensus alone is therefore not evidence of collective learning. The prospective

experiment shows that this contraction depends on the communication format, but retained dispersion may reflect useful minority information, noise, or poor forecasts. The crux-ledger result supports a communication-design hypothesis, though it bundles claim summarization, deduplication, and attribution removal without isolating their separate effects. Follow-up evaluation should test resolved outcomes, calibration, and accuracy-adjusted links between prompt diversity and aggregation gains. Transfer to resource-constrained forecasting with local models remains an open question; such comparisons should assess predictive quality and inference cost jointly. Further, future work could test whether hierarchical architectures, i.e., aggregating small homogeneous sub-ensembles within a broader heterogeneous model committee, can balance intra-model variance reduction with genuine inter-model error diversity.

6 Conclusion

In this same-model, shared-evidence setting, forecasts show strong residual dependence, and diversity-aware prompt selection does not detectably outperform one random ensemble. Standard deliberation contracts disagreement without a detectable aggregate accuracy improvement. A probability-blind crux ledger preserves substantially more dispersion, but whether it preserves useful predictive information remains unresolved. Forecast diversity and consensus should therefore be evaluated against outcomes, not treated as evidence of collective intelligence on their own.

7 Limitations

The 189 resolved snapshots cover 165 markets in limited domains. Market-clustered bootstrapping does not assume independent snapshots within a market, but treats market clusters as independent. Related contracts, shared events, and common news shocks may violate that assumption and make intervals too narrow. Event-clustered or one-contract-per-event sensitivity analyses are needed; these have not been performed here. The sample limits precision and generalizability, particularly for null findings.

Holding model and evidence fixed isolates prompt variation, but does not establish transfer to open-weight models, other providers, or heterogeneous ensembles. Single responses do not separate prompt effects from sampling variation. Effective rank and correlations are sample-dependent descriptors, not direct measures of unique predictive value. One random comparator does not establish selection-method equivalence. We do not separately assess calibration or overconfidence. The prospective experiment uses a different model and outcomes unresolved at evaluation, so it cannot establish accuracy, calibration, or decision-quality benefits. Two of its communication conditions are exploratory.

8 Ethical Considerations

We use public market and news data, recruit no participants, collect no private data, and place no trades. The forecasts may be inaccurate and should not be treated as financial advice; generated briefs and rationales may also reproduce factual errors or biases. Because event-level forecasts could influence thin unresolved markets, such artifacts should be withheld until resolution.

References

- Matthew Aitchison, Scott Jeen, Toby Shevlane, and Ben Day. 2026. [Diversity is the strength of the AI crowd](#). *Preprint*, arXiv:2606.29661.
- Pavel Atanasov, Phillip Rescober, Eric Stone, Samuel A. Swift, Emile Servan-Schreiber, Philip Tetlock, Lyle Ungar, and Barbara Mellers. 2017. [Distilling the wisdom of crowds: Prediction markets vs. prediction polls](#). *Management Science*, 63(3):691–706.
- Nuo Chen, Yicheng Tong, Yuzhe Yang, Yufei He, Xueyi Zhang, Qingyun Zou, Qian Wang, and Bingsheng He. 2026. [Diversity collapse in multi-agent LLM systems: Structural coupling and collective failure in open-ended idea generation](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 251–306. Association for Computational Linguistics.
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. [Improving factuality and reasoning in language models through multiagent debate](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 11733–11763. PMLR.
- Tobias Groot and Matias Valdenegro-Toro. 2024. [Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models](#). In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pages 145–171, Mexico City, Mexico. Association for Computational Linguistics.
- Wenyang Hu, Gregory Kang Ruey Lau, Diwen Liu, Jizhuo Chen, See-Kiong Ng, and Bryan Kian Hsiang Low. 2025. [DIPPER: Diversity in prompts for producing large language model ensembles in reasoning tasks](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35558–35572, Suzhou, China. Association for Computational Linguistics.
- Giulia Iadisernia and Carolina Camassa. 2025. [Prompting for policy: Forecasting macroeconomic scenarios with synthetic LLM personas](#). In *Proceedings of the 6th ACM International Conference on AI in Finance*, pages 335–343.
- Anders Krogh and Jesper Vedelsby. 1994. [Neural network ensembles, cross validation, and active learning](#). In *Advances in Neural Information Processing Systems*, volume 7, pages 231–238. MIT Press.
- Yuante Li, Yicheng Tao, Kate Zhang, Taozhi Wang, Gefei Gu, and Yaxin Zhou. 2026. [Diverse evidence, better forecasts: Multi-agent deliberation under information asymmetry](#). *Preprint*, arXiv:2607.01661.
- Jan Lorenz, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. 2011. [How social influence can undermine the wisdom of crowd effect](#). *Proceedings of the National Academy of Sciences*, 108(22):9020–9025.
- Dipkumar Patel. 2026. [Representational collapse in multi-agent LLM committees: Measurement and diversity-aware consensus](#). *Preprint*, arXiv:2604.03809.
- Olivier Roy and Martin Vetterli. 2007. [The effective rank: A measure of effective dimensionality](#). In *15th European Signal Processing Conference*, pages 606–610, Poznan, Poland.
- Paul Schneider and Amalie Schramm. 2025. [The wisdom of deliberating AI crowds: Does deliberation improve LLM-based forecasting?](#) *Preprint*, arXiv:2512.22625.

Philipp Schoenegger, Indre Tuminauskaite, Peter S. Park, Rafael Valdece Sousa Bastos, and Philip E. Tetlock. 2024. [Wisdom of the silicon crowd: LLM ensemble prediction capabilities rival human crowd accuracy](#). *Science Advances*, 10(45):eadp1528.

Andries Petrus Smit, Nathan Grinsztajn, Paul Duckworth, Thomas D. Barrett, and Arnau Pretorius. 2024. [Should we be going MAD? a look at multi-agent debate strategies for LLMs](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 45883–45905. PMLR.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. [Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. [Self-consistency improves chain of thought reasoning in language models](#). In *The Eleventh International Conference on Learning Representations*.

Zengqing Wu and Takayuki Ito. 2025. [The hidden strength of disagreement: Unraveling the consensus-diversity tradeoff in adaptive multi-agent systems](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15277–15297. Association for Computational Linguistics.

Yingxuan Yang, Chengrui Qu, Muning Wen, Laixi Shi, Ying Wen, Weinan Zhang, Adam Wierman, and Shangding Gu. 2026. [Understanding agent scaling in LLM-based multi-agent systems via diversity](#). Preprint, arXiv:2602.03794.

A Forecasting Performance

Table 1 reports the forecasting comparisons discussed in the main text. Confidence intervals and paired differences are bootstrapped by market. These comparisons provide forecasting context but are not the primary object of the study.

B Persona Prompt Templates

Each candidate prompt combines one of the three fixed persona prefixes below with a distinct generated reasoning-strategy instruction. The question, resolution rules, and output instructions are otherwise identical across prompts. The full set of 285 candidate prompts and the 30 selected prompts is included in the public repository.

Outside-view analyst.

You are an outside-view analyst. Ground your probability estimate in the historical base rate for this type of event and the relevant reference class. Treat this case as one draw from a distribution of similar past events, and resist adjusting away from the base rate unless you have specific, well-evidenced reasons to do so.

Inside-view analyst.

You are an inside-view analyst. Focus on the specific causal mechanisms, actors, and conditions driving this particular situation. Build your probability estimate from the details of this case rather than historical averages, and weight recent developments and case-specific evidence heavily.

Superforecaster.

You are a superforecaster. Integrate multiple perspectives before committing to a probability. Actively seek out reasons your initial estimate might be wrong. Express calibrated uncertainty—distinguish clearly between what you know and what you are uncertain about—and update your estimate when the evidence warrants it.

Shared prompt suffix. Each persona prefix is followed by the generated instruction and the following shared fields:

```
[Generated reasoning-strategy instruction]
Question: {question_text}
Resolution rules: {resolution_rules}
Return JSON: {"probability": 0.0,
              "reasoning": "one short sentence"}
```

C Interpreting Aggregation and Diversity

Finite-poll subset averaging. For a fixed forecast panel of $N > 1$ agents, let $D_m = N^{-1} \sum_i (p_{im} - \bar{p}_m)^2$ and let S be a uniformly sampled subset of k distinct agents. The subset mean $\bar{p}_{S,m}$ is unbiased for $\bar{p}_m$, and its sampling variance is $\text{Var}_S(\bar{p}_{S,m}) = (N - k)D_m/[k(N - 1)]$. Therefore,

$$\mathbb{E}_S[(\bar{p}_{S,m} - y_m)^2] = (\bar{p}_m - y_m)^2 + \frac{N - k}{k(N - 1)} D_m. \quad (1)$$

Averaging over the same snapshots and weights gives $\mathbb{E}[\text{BS}_k] = \text{BS}_N + (N - k)D/[k(N - 1)]$, where D is mean within-panel variance. Since $\mathbb{E}[\text{BS}_1] - \text{BS}_N = D$, the fraction of the full-poll gain attained at size k is

$$\frac{\mathbb{E}[\text{BS}_1] - \mathbb{E}[\text{BS}_k]}{\mathbb{E}[\text{BS}_1] - \text{BS}_N} = \frac{N(k - 1)}{k(N - 1)}, \quad (2)$$

provided $D > 0$. At $N = 30$, the shares are 82.8% for $k = 5$ and 93.1% for $k = 10$, regardless of independence. Thus, the score levels and absolute gains are empirical, whereas these normalized shares are algebraic.

Forecast	Brier score	95% CI	Δ vs. DIPPER	95% CI
DIPPER-selected poll	0.2108	[0.1788, 0.2468]	—	—
Random-prompt poll	0.2118	[0.1799, 0.2477]	+0.0010	[−0.0048, 0.0063]
Team poll	0.2096	[0.1761, 0.2457]	−0.0013	[−0.0093, 0.0075]
Prediction market	0.1472	[0.1178, 0.1809]	−0.0637	[−0.0964, −0.0328]
<i>Always-0.5 baseline</i>	0.2500	—	—	—
<i>Uniform-random baseline</i>	0.3333	—	—	—

Table 1: Forecasting performance across 189 market snapshots. Lower Brier scores are better; negative differences indicate improvement over the DIPPER-selected poll. Confidence intervals and paired differences are bootstrapped by market. Shaded rows denote uninformative reference baselines. The uniform-random score is the expected Brier score for probabilities drawn from $U[0, 1]$.

Forecast differences and predictive value. Let $R_i = \mathbb{E}_m[(p_{im} - y_m)^2]$, with $\mathbb{E}_m$ denoting the empirical snapshot average. For the equal-weight mean of agents i and j ,

$$\frac{R_i + R_j}{2} - \text{BS}_{ij} = \frac{1}{4} \mathbb{E}_m[(p_{im} - p_{jm})^2]. \quad (3)$$

Thus, squared forecast distance directly determines the gain relative to the pair’s mean individual error. It does not establish a better absolute pair score if the individual errors are also larger. Equivalently, setting $e_{im} = p_{im} - y_m$ gives

$$\text{BS}_{ij} = \frac{R_i + R_j + 2\mathbb{E}_m[e_{im}e_{jm}]}{4}. \quad (4)$$

Residual correlation alone omits error means and scales. The reported semantic-distance correlations therefore establish association with forecast variation, not a causal or held-out benefit from selecting semantically distant prompts. No new pairwise performance analysis is reported here; such an analysis should account for constituent accuracy and dependence among pairs sharing prompts or markets.

Covariance-spectrum effective rank. For the forecast covariance matrix, let q_j denote its normalized eigenvalues. We report the entropy-based measure $r_{\text{eff}} = \exp(-\sum_j q_j \log q_j)$, taking $0 \log 0 = 0$ (Roy and Vetterli, 2007). This covariance-spectrum version is near one when variation lies mainly along one direction and reaches N when all eigenvalues are equal and positive. It describes how forecast variance is spread across directions, not the number of independent agents or the usefulness of their predictions.